Friday, Aug 31st 2018
NEURO BOOT CAMP – DAY 5

Population–level analysis

Problem 1: Information of a symbol string

Consider an alphabet of 8 symbols whose probabilities are as follows:

A	B	C	D	E	F	G	H
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{128}$

• (a) If someone has selected one of these symbols and you need to discover which symbol it is by asking yes/no questions that will be truthfully answered, what would be the most efficient sequence of such questions that you could ask in order to discover the selected symbol?

• (b) By what principle can you claim that each of your proposed questions is maximally informative?

• (c) On average, how many such questions will need to be asked before the selected symbol is discovered?

• (d) What is the entropy of the above symbol set?

• (e) Construct a uniquely decodable binary code for the symbol set. Relate the bits in your prefix code to the yes/no questions that you proposed in (a).

Problem 2: Entropy definitions

Let X and Y represent random variables with associated probability distributions $p(x)$ and $p(y)$, respectively. They are not independent. Their conditional probability distributions are $p(x|y)$ and $p(y|x)$, and their joint probability distribution is $p(x, y)$.

• (a) What is the marginal entropy $H(X)$ of variable X, and what is the mutual information of X with itself?

• (b) In terms of the probability distributions, what are the conditional entropies $H(X|Y)$ and $H(Y|X)$?

• (c) What is the joint entropy $H(X, Y)$, and what would it be if the random variables X and Y were independent?

• (d) Give an alternative expression for $H(Y) - H(Y|X)$ in terms of the joint entropy and both marginal entropies.

• (e) What is the mutual information $I(X; Y)$?

Problem 3: Correlation and covariance in neuronal populations

• (a) If $Y = 2X - 3Z$ where X and Z are independent random variables with $Var(X) = 3$, $Var(Z) = 2$, determine the correlation of X and Y.

Hint: for any random variables S, T, R and any constants a, b, write down a simplified expression for both $Cov(S, aT + bR)$ and $Cov(S, aT)$

• (b) In today's folder you will find 4 different datasets. Load the large matrix from each one of the three imaging datasets (mouse retrosplenial cortex, hippocampus CA1 and primary visual cortex). Each one of them contains both a large matrix of neurons vs time and another variable which is correlated with the neural activity. For now, lets set aside the other variables and have a look at the neural activity matrices alone. Unless specified otherwise, you may choose to work with the the full dataset or with a subset of neurons from each one.

Play around with the matrices – is there any difference is the format of the signal between them? Compute whatever you think is interesting (feel free to take your time). Is there anything you can say about the structure of the neural activity, what is responds to, or the behavior of the animal?

• (c) Lets try to investigate a little more systematically. For each one of the three matrices of neural activity:
1. Compute the variance of 3 randomly chosen neurons;
2. Compute the covariance of 3 randomly chosen pairs of neurons;
3. Compute the correlation coefficient of the same 3 pairs you computed the covari-

ance for;
4. Write down the 9 values from either (2) or (3) on paper (scramble their order) and show to a the friend sitting to your right. Can she/he tell what values belong to the same dataset?

• (d) Choose one dataset.
1. Visualize the correlation matrix of all pairs of neurons. Does it look right? (try to subtract the diagonal for better visualization)
2. Visualize the covariance matrix of all pairs of neurons. Does it look right? (try to subtract the diagonal for better visualization)

• (e) Plot the probability density function of both covariance and correlation values (separately). Is their shape similar? Are you surprised by it? What do you think influences the shape of these distributions?

• (f) Create a binarized version of the dataset you chose and repeat (d). What are the differences? Are we going to loose much by treating the signal as a binary time series?

• (g) Drawing on the expression for Pearson correlation coefficient that you are familiar with for a pair of variables, write down an expression for the correlation between 3 neurons. Choose a subset of neurons and plot the pdf for all 3-neuron correlation values of the set. Does it look similar to the pairwise correlations distributions you plotted before?

• (h) Determine what is the relevant quantity to compute in order to tell how sparse is the neural activity. Are the different brain regions very different in that aspect?

Problem 4: Information about an external variable

A significant fraction of rodent hippocampal neurons can be classified as "place cells". These are neurons that fire only when the animal visits a particular location in its environment. As the animal moves through the "place field" of a specific place cell, the firing rate of that neuron increases. As the animal leaves the location, the firing stops. The following steps will help you determine which neuron is a place cell and which one is not

Load the hippocampus CA1 imaging dataset and the associated vector of position of the animal at every time point. Visualize the neural activity and the position;

examine the repetitions of trials.

• (a) Write down the expression for mutual information between an individual neuron's activity and position. Verify that it is the same expression as in *Skaggs et al (1993)*

• (b) Write down the steps you need to take in order to compute (a) numerically for one of the neurons in the dataset. This is a somewhat flexible measure that depends on personal choices, so think carefully about how to binarize the data, how to bin the position, etc.

• (c) About $30\% - 50\%$ of neurons in a population of CA1 cells are place cells. Use this this fact to determine an appropriate threshold for the amount of information about position which a neuron has to possess to qualify as a place cell. How many neurons in the dataset are place cells?

• (d) Choose two place cells with non-overlapping fields. What is the time lag T that optimizes the cross correlation between them? (In case you would like to work in seconds, the acquisition rate of the data is 30 Hz)

• (e) What is the average correlation between place cells? and between non place cells? Plot the correlation matrix for the place cell and the non place cells separately. Are their values in the same range? Is that to be expected?

Problem 5: Principal component analysis from scratch

Principal Component Analysis, or PCA for short, is a method for reducing the dimensionality of data. It can be thought of as a projection method where data with m-columns (features) is projected into a subspace with m or fewer columns, whilst retaining the essence of the original data. The PCA method can be described and implemented using the tools of linear algebra. PCA is an operation applied to a dataset, represented by an n x m matrix A that results in a projection of A which we will call B. Lets walk through the steps of this operation.

1. First we calculate the mean values of each column.

2. Center the values in each column by subtracting the mean column value.

3. Center the values in each column by subtracting the mean column value. The next step is to calculate the covariance matrix of the centered matrix C.

Correlation is a normalized measure of the amount and direction (positive or negative) that two columns change together. Covariance is a generalized and unnormalized version of correlation across multiple columns. A covariance matrix is a calculation of covariance of a given matrix with covariance scores for every column with every other column, including itself.

4. Finally, we calculate the eigen-decomposition of the covariance matrix V. This results in a list of eigenvalues and a list of eigenvectors.

The eigenvectors represent the directions or components for the reduced subspace of B, whereas the eigenvalues represent the magnitudes for the directions.

The eigenvectors can be sorted by the eigenvalues in descending order to provide a ranking of the components or axes of the new subspace for A. If all eigenvalues have a similar value, then we know that the existing representation may already be reasonably compressed or dense and that the projection may offer little.

If there are eigenvalues close to zero, they represent components or axes of B that may be discarded.

5. A total of m or less components must be selected to comprise the chosen subspace. Ideally, we would select k eigenvectors, called principal components, that have the k largest eigenvalues.

Other matrix decomposition methods can be used such as Singular-Value Decomposition, or SVD. As such, generally the values are referred to as singular values and the vectors of the subspace are referred to as principal components.

6. Project the data into subspace via matrix multiplication. (Multiply the transpose of the chosen principal components with the original data)

Problem 6 (optional): Gaussian maximizes entropy

Recall the definition of entropy as a measure for the amount of randomness in a variable. For a random variable x: $H(x) = \Sigma_i p(x_i) log(p(x_i))$.

Show that entropy is maximized if p is uniform.